# 1 Training Algorithm

**Algorithm 1** Training for Image-to-Video network

**Input:** Untrimmed video set $\{V_i\}_{i=1}^{N}$, Video-level labels $\{\boldsymbol{y}_i\}_{i=1}^{N}$
**Output:** Updated I2V model

1: Initial network with ImageNet pretrained model
2: **for** Each epoch **do**
3: Sample several frames $\{T_{k,i}\}_{k=1}^{K}$ from Videoc $V_i$
4: Feed all sampled frames to I2V netweork to get the outputs $\{\boldsymbol{z}_{T_{k,i}}\}_{k=1}^{K}$
5: Do average consensus among outputs of all frames to get video-level prediction $\boldsymbol{z}_i = \frac{1}{K}\sum_{k=1}^{K}\boldsymbol{z}_{T_{k,i}}$
6: Back propagate and update model
7: **end for**

**Algorithm 2** Training for Video-to-Proposal network

**Input:** Action proposal set $\{\{(t^q_{start,i}, t^q_{end,i})\}_{q=1}^{Q}\}_{i=1}^{N}$, Video-level labels $\{\boldsymbol{y}_i\}_{i=1}^{N}$
**Output:** Updated V2P model

1: Initial network with pretrained I2V model
2: **for** Each epoch **do**
3: Feed all action proposals $\{(t^q_{start,i}, t^q_{end,i})\}_{q=1}^{Q}$ of $V_i$ to V2P network to get the output $\{\boldsymbol{r_q}\}_{q=1}^{Q}$
4: Do maximun consensus $v_i^j = \max(r_1^j, r_2^j, \ldots, r_n^j)$ among outputs of all proposals to get video-level prediction $\hat{\boldsymbol{p}}_{\boldsymbol{i}} = softmax(\boldsymbol{v}_i)$
5: Back propagate and update model
6: **end for**

# 2 Testing Algorithm

**Algorithm 3** Action localization

**Input:** Action proposals $\{\{(t^q_{start,i}, t^q_{end,i})\}_{q=1}^{Q}\}_{i=1}^{N}$, threshold $\theta$
**Output:** Detection results

1: Feed all proposals to trained V2P network and get scores for each proposal
2: Implement NMS for each class among all proposals
3: Select propsals with score higer than a threshold $\theta$ as the final detection results